Prosodic marking, pitch and intensity in spontaneous lexical self-repair in Dutch

Leendert Plug, Paul Carter

Abstract

This paper presents results of a phonetic analysis of instances of lexical self-repair drawn from a corpus of spontaneous Dutch speech. The analysis addresses questions concerning the phonetic details of prosodic marking in self-repair and its conditioning factors. In particular, it examines the relevance of semantic, temporal and frequency-related factors in modelling f0 and intensity measures and auditory judgements of whether repairs are prosodically marked. It addresses the extent to which observations made in studies using experimentally-elicited speech can be expected to generalise to repairs drawn from uncontrolled spontaneous speech. The results suggest that prosodic marking is rare in spontaneous lexical self-repair, and that semantic, temporal and frequency factors play a limited role only in conditioning speakers' choices for or against prosodic marking, although several weak tendencies can be observed.

INTRODUCTION

In this paper we report on a phonetic analysis of instances of lexical self-repair such as *I'm going on Thursd- Friday*, in which one lexical choice — here *Thursday* — is rejected in favour of another — here *Friday*. While a good deal is known about the various types of disfluency involved in the initiation of self-repair [see e.g. Nakatani and Hirschberg 1994, Shriberg 2001, Jasperson 2002, Benkenstein and Simpson 2003], relatively few studies have addressed the question of how the phonetics of the second — preferred — lexical item compare to those of the first — rejected — one. The main references on this question remain Cutler [1983] and Levelt and Cutler [1983], who establish the notion of 'prosodic marking' in self-repair.

Levelt and Cutler on prosodic marking

On the basis of analysis of an unspecified number of spontaneous speech error repairs, Cutler [1983] concludes that in producing a self-repair, a speaker has a choice between prosodically 'marking' the repair, and leaving it 'unmarked'. She describes an 'unmarked' repair as one in which the pitch, intensity and speaking rate of the preferred lexical item — henceforth the *repair item* — are not noticeably different from those of the rejected lexical item — henceforth the *reparandum item*. A 'marked' repair, on the other hand, 'is distinguished by a

quite different prosodic shape from that of the original utterance' [Cutler 1983: 81]. By leaving a repair unmarked, the speaker 'minimises the disruptive effect of the error on the utterance as a whole', while marking assigns 'salience' to the correction [Cutler 1983: 80].

On the basis of an independent study of Dutch task-oriented speech, Levelt and Cutler [1983] claim that the speaker's choice for or against prosodic marking is constrained to some extent by the semantics of the repair that is being produced. First, like Levelt [1983], Levelt and Cutler distinguish between 'error repairs', in which a factual or linguistic error is corrected, and 'appropriateness repairs', in which the problem with the initial lexical choice is one of felicity rather than error. The example of *Thursd- Friday* above illustrates error repair: *Thursday* and *Friday* have mutually exclusive denotations, so if one is factually accurate the other cannot be. An example of appropriateness repair would be *I saw that guy- uh, man yesterday,* where *guy* and *man* have the same referent, but the latter is — presumably — deemed more appropriate by the speaker than the former, given the pragmatic context. Levelt and Cutler observe that while a majority of error repairs in their data are perceivable as prosodically marked, a majority of appropriateness repairs are perceivable as unmarked.

Second, Levelt and Cutler [1983] claim that for error repairs, an additional factor constraining speakers' choice for or against prosodic marking is the size of the semantic field to which the reparandum and repair items belong. When this is finite, as in the case of days of the week, prosodic marking is more likely when the set is smaller. Levelt and Cutler observe the effect when they compare corrections of colour terms, of which 11 are relevant in the task their participants are performing, and directions, of which 4 are relevant: while about half of colour corrections are produced with noticeable prosodic marking, 72% of direction corrections are.

Levelt and Cutler's work raises a number of interesting questions which so far have not been addressed in detail. One concerns the phonetic details of prosodic marking in self-repair. Having defined a prosodically marked repair as one characterised by 'a noticeable increase or decrease in pitch, in amplitude, or in relative duration', Levelt and Cutler [1983: 206] make no attempt to describe the instances they consider marked in terms of their pitch, amplitude and duration characteristics. Cutler [1983: 80–81] indicates that 'typically', a marked repair 'is uttered on a higher pitch and with greater intensity than the erroneous material', but some marked repairs are perceivable as such 'by being uttered on a noticeably lower pitch'. Later on, she suggests that marking can be realised 'in several different ways — by longer relative duration, noticeably higher or lower pitch, noticeably higher or lower amplitude, or a combination of pitch, amplitude and durational effects' [Cutler 1983: 84].

However, although she refers to instrumental analysis in the discussion of selected instances, she does not present quantitative evidence to back up her generalisations. As a result, we are left to wonder to what extent speakers manipulate pitch, intensity and speaking rate independently in prosodic marking in self-repair, and how many types of marking are likely to be attested in any sizeable corpus of self-repairs.

Researchers working on sound patterns in spontaneous conversation tend to emphasize the importance of detailed analysis of the 'clusters' of phonetic features that give rise to auditory impressions of 'emphasis', 'phonetic upgrading' and so on, guided by the principle that there is no *a priori* way of predicting how these clusters will be constituted in any given context, and the hypothesis that different phonetic implementations may serve different communicative purposes [Local 2003, Local and Walker 2005, Selting 2010: 27]. From this point of view, the definitions of prosodically marked repairs provided by Levelt and Cutler [1983] and Cutler [1983] are intriguing, and warrant investigation that involves both auditory and acoustic analysis.

A second question concerns the factors conditioning prosodic marking in self-repair — in particular the extent to which they generalise beyond Levelt and Cutler's [1983] corpus of task-oriented speech. It would seem plausible that in task-oriented dialogue, error corrections in some sense carry more weight than appropriateness repairs, since the success of the task depends crucially on the correctness — in particular the factual accuracy — of the participants' instructions to each other as they perform the task. By contrast, the success of the task does not depend crucially on whether participants choose the pragmatically most felicitous way of formulating their utterances. However, in unrestricted spontaneous dialogue this may well be different. It does not, in principle, seem difficult to conceive of discourse scenarios in which an appropriateness repair carries more weight than a correction of factual accuracy or linguistic well-formedness: for example, an inappropriately phrased reference to a person familiar to both conversation partners is likely to have an observable impact on subsequent turns in the interaction; a topically peripheral error of fact or an isolated instance of ungrammatically is not. We might wonder, then, whether the effect of the error *versus* appropriateness dichotomy described by Levelt and Cutler [1983] will be attested in a corpus of self-repairs drawn from genuinely spontaneous speech. Moreover, we might wonder whether the effect of semantic field size will be reflected in effects of frequency-related measures. We will return to the latter below.

Subsequent studies

Insofar as these questions have been addressed in subsequent studies, these have failed to produce strong evidence for the generalisability of Cutler's [1983] and Levelt and Cutler's [1983] findings. With reference to the phonetics of prosodic marking in self-repair, Howell and Young [1991] report a weak tendency for lexical repairs sampled from the Survey of English Usage to be accompanied by a rise in pitch and intensity between the two lexical items involved; Nakatani and Hirschberg [1994] report a similar result for repairs sampled from the American English ARPA Air Travel Information System corpus. Nakatani and Hirschberg [1994: 1611] emphasize that '[w]hile we find small but significant changes in two correlates of intonational prominence, the distributions of change in *f0* and energy for our data are unimodal': in other words, while there is some evidence that repairs may be prosodically marked by a rise in pitch and intensity, there is little evidence that marking is achieved through a noticeable fall along these parameters, as suggested by Cutler [1983], with any frequency. Hokkanen [2001] and Cole et al. [2005] come to the same conclusion with respect to pitch, on the basis of Finnish and American English data respectively.

With reference to conditions on prosodic marking, none of the studies mentioned above includes a semantic analysis of the repairs in their data. As far as we know, the only attempt to replicate Levelt and Cutler's [1983] analysis is made by Plug [2011], who investigates the temporal organisation of a small collection of spontaneous self-repairs sampled from Dutch spontaneous speech. Plug reports a predominance of temporal compression across the repair item relative to the reparandum item. He finds no significant effect on relative repair tempo of the 'error' versus 'appropriateness' dichotomy, and no significant effect of the difference in word frequency between the two lexical items involved in the repair.

In addition, there is reason to doubt the generalisability of another of Cutler's [1983] and Levelt and Cutler's [1983] findings, which we have not mentioned so far. In addition to reporting the effects of repair semantics described above, Levelt and Cutler [1983: 211] report no significant effect on the likelihood of prosodic marking of what they call the 'interruption-and-restart structure of the repair'. In particular, repairs in which the reparandum item is interrupted prematurely are not more or less likely to be prosodically marked than repairs in which the reparandum item is completed — or even followed by a pause or additional lexical material — before the onset of repair. Similarly, Cutler [1983: 81] describes the choice between marking and not marking a repair as 'apparently orthogonal to the time course of error detection and correction'. This is challenged by the observation by

Levelt [1989: 481] and Brédart [1991] that error repairs are more likely than appropriateness repairs to involve an early interruption of the reparandum item. Moreover, Nooteboom [2010] has reported consistent prosodic differences between phonological error repairs following early and late interruptions of the erroneous utterance. Nooteboom observes that repairs in which the interruption comes very early, as in *sa … fat soap*, tend to be associated with a high pitch and intensity prominence on the first vowel of the repair. Instances in which the erroneous word is completed before the onset of repair tend to be associated with a low pitch and intensity prominence on the first vowel. This warrants a reconsideration of the relationship between the temporal make-up of lexical repairs and their prosodic characteristics.

Finally, Kapatsinski [2010] has shown that in American English, there is a predictable relationship between the temporal make-up of lexical repairs and the frequency of the words involved in lexical repair, such that high-frequency reparandum items are less likely to be cut off prematurely prior to repair than low-frequency items. With frequency measures possibly capturing some of the effect of semantic field size reported by Levelt and Cutler [1983] and co-varying with temporal measures, we might expect to find interesting interactions between semantic, temporal and frequency-related factors in accounting for the prosody of lexical self-repair.

This study

In this paper we report on an attempt to model pitch and intensity characteristics of a collection of lexical repairs sampled from the Corpus Spoken Dutch [Oostdijk 2002]. We derive the characteristics from auditory judgements of prosodic marking, following Cutler [1983] and Levelt and Cutler [1983], as well as acoustic measurements, following Nakatani and Hirschberg [1994], Nooteboom [2010] and others. We explore the relationship between the auditory judgements and measurements in the light of Levelt and Cutler's [1983] suggestion that the perception of prosodic marking in self-repair can be triggered by a variety of prosodic relationships between reparandum and repair. Moreover, we evaluate the role of semantic, temporal and frequency-related factors in accounting for both.

MATERIAL AND METHODS

Data selection

The data set for this paper comprises 216 instances of lexical repair extracted from four sub-corpora of the Spoken Dutch Corpus [Oostdijk 2002], containing spontaneous face-to-face conversations, interviews with teachers of Dutch, broadcast interviews, discussions and debates, and non-broadcast interviews, discussions and debates. We searched for instances of speech which were coded as interrupted and for a selection of lexical items that may function as 'editing terms' in the context of repair [Levelt 1983], including *of* 'or', *nee* 'no' and *eigenlijk* 'actually' — as well as performing a number of additional, unsystematic data trawls. We discarded a considerable number of potential instances because of poor audio quality or overlapping speech. We left aside instances in which the reparandum item was left incomplete and either no reasonable guess could be made as to its identity, or several candidate identities presented themselves. This selection was done by the first author in the first instance, and was later verified by the independent linguist who assisted in the semantic classification of the repairs, as described below. We also left aside clause-initial and clause-final repairs, to minimise the effect of boundary contours — in particular clause-final rises and falls — on our prosodic measurements.

(1) contains representative examples from our data set. The reparandum and repair items are in bold. The examples in (1) illustrate that some cases the reparandum item is cut off prematurely, as in (a), (b), (c) and (g), and in others it is completed, as in (d) to (f). In some cases, lexical material preceding the reparandum item is repeated in the repair, as in (a), (d), (e) and (g); and in some cases, the repair is initiated by an editing term such as *of* in (d) and (f) or the hesitation marker *uh* in (g). We will return to some of these characteristics below.

(1) a. met de **au-** met de **bus** ('by ca- by bus')
b. als er met tekst **gebrui- gewerkt** wordt ('when one use- works with text')
c. de **koelka- koelcel** ('the refrigera- cold store')
d. die **drie** da- of die **twee** dagen ('those three day- or those two days')
e. een **leuke** k- een **mooie** keuken ('a nice k- a beautiful kitchen')
f. **een** telefoon- of **mijn** telefoonnummer opschrijven ('write down a phone- or my phone number')
g. in de **computerwe-** uh in de **bankwereld** ('in the world of compu- er of banking')

Segmentation and acoustic analysis

We segmented all instances of repair in PRAAT [Boersma and Weenink 2010]. We placed boundaries at the starts and ends of the two crucial lexical items involved in the repair, and delimited the vowel portions within these intervals, following the segmentation criteria set out by Rietveld and Van Heuven [1997]. The number of vowel portions ranges from 1 to 5 for the reparandum item, and from 1 to 7 for the repair item. Figure 1 illustrates the segmentation.

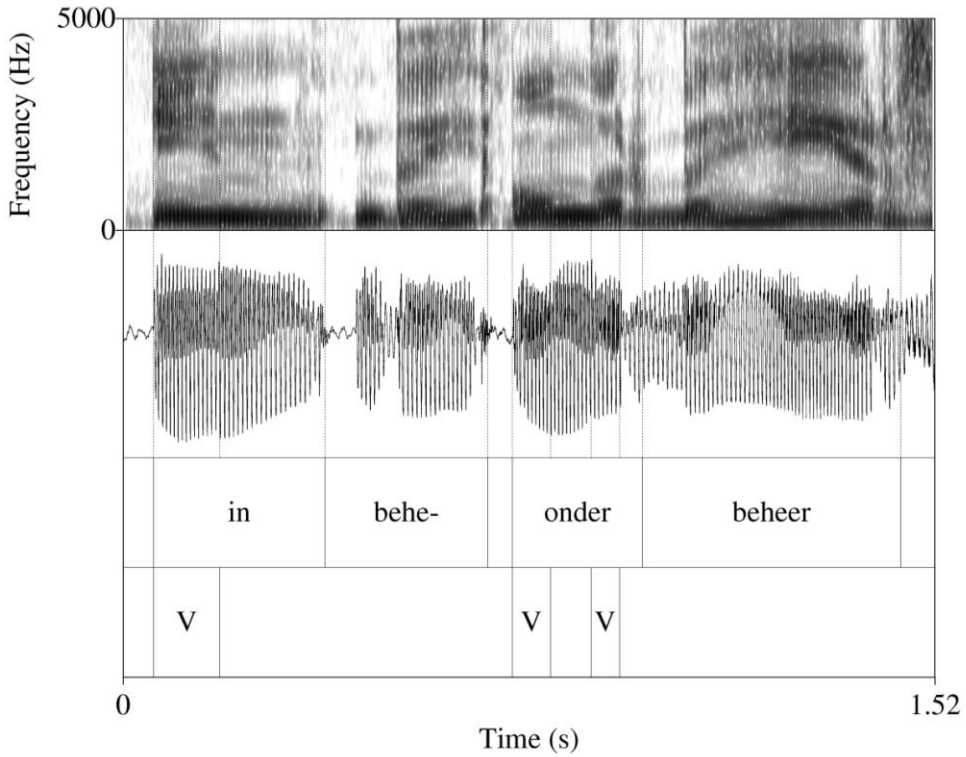


Figure 1. Segmented spectrogram and waveform for the repair in (1e). f0 and intensity measurements were taken in the three vowel portions labelled 'V': one for the reparandum item *in* /ɪn/, and two for the repair item *onder* /ɔndər/.

We measured f0 (in Hertz) and intensity (in decibels) at every millisecond across the segmented vowel portions, and log-transformed f0 values. We then calculated maximum, median and mean values, and calculated corresponding delta values by subtracting the value derived from the reparandum item from that derived from the repair item. This yields a measure of the prosodic difference between the two lexical items involved in the repair, and introduces a degree of speaker normalisation. To illustrate, a positive value for f0 maximum means the instance has a repair item whose highest f0 value is above that of the highest f0 value associated with the reparandum item.

Prosodic marking

Following Levelt and Cutler [1983], we classified all instances as prosodically marked or unmarked based on auditory analysis. The crucial question in each case was whether the repair item sounds particularly salient because of its pitch or loudness, or a combination of the two, relative to the reparandum item. Unlike Levelt and Cutler [1983], we allowed for the intermediate classification of 'possibly marked'.

The classification was done by two raters: the first author, who is a native speaker of Dutch and an academic phonetician, and a Dutch linguist with a research specialisation in pragmatics and discourse studies. The latter had no particular expectations as to which types of repair should or should not be marked. The two raters independently classified all instances by listening to the repair in the context of one or two preceding and following words. They reached the same judgement in 182 cases (84%): 'marked' in 31, 'unmarked' in 147 and 'possibly marked' in four. Of the 34 instances for which the raters proposed a different classification, 21 involved one rater proposing 'possibly marked' and the other either 'marked' or 'unmarked'. In order not to underestimate the proportion of instances with some degree of prosodic marking, we coded a combination of 'marked' and 'possibly marked' (7 instances) as 'marked', and a combination of 'unmarked' and 'possibly marked' (14 instances) as 'possibly marked'. The remaining 13 instances for which one rater proposed 'marked' and the other 'unmarked' were reconsidered by the rater who had proposed 'unmarked'. In all cases this rater accepted a coding of either 'possibly marked' or 'marked'. In the final coding, 43 instances (20%) are classified as prosodically marked, 24 (11%) as possibly marked and 149 (69%) as prosodically unmarked. The percentage of clearly marked instances is low compared with the marking percentages reported for lexical repair by Cutler [1983] and Levelt and Cutler [1983] — 38% and 45% respectively — even if our 'possibly marked' instances are counted as marked for comparison. We will return to this observation below.

In what follows, we will refer to the marking classification by the name we gave to this variable in our analysis, *Prosodic marking.*

Repair semantics

In order to assess the predictive value of Levelt and Cutler's [1983] 'error' versus 'appropriateness' dichotomy, we classified all instances as error or appropriateness repair using the criteria set out by Levelt [1983] and, more recently, Kormos [1999]. (We will refer to the resulting variable as *Repair type* in what follows.) Generally, instances in which the

denotations of the two lexical items are mutually exclusive, as in (1a), (1d) and (1g) above, or in which the first lexical choice result in an ill-formed collocation, as in (1b), can be considered error repairs. Instances in which the denotations of the two lexical items are highly similar, as in (1c) and (1e), can be considered appropriateness repairs. In these cases, the first lexical choice is treated as ill-judged by the speaker, but is not factually or linguistically erroneous. Instances in which the second lexical item can be seen as more specific than the first, as in (1f), can also be considered appropriateness repairs.

The classification was done by the same two raters who did the auditory judgements. The two classifications were completed almost a year apart, so that the probability that the independent Dutch linguist was influenced by one when doing the other is minimal — particularly given that she was neither familiar with the details of Levelt and Cutler's [1983] findings nor with the specific aims of the current study. The data set considered contained 222 instances. As indicated above, instances of repair with an incomplete reparandum item in which the identity of the item could not be established with reasonable certainty by the first author were not included in the data set. The second rater first verified that the first author's interpretations of the incomplete reparandum items that were included in the data set were reasonable. Unfortunately, some previous studies have left repair semantics aside on the grounds that its analysis 'would have involved far too many guesses' [Howell and Young 1991: 741], or restricted semantic analysis to repairs with completed reparandum items [Kapatsinski 2010: 90]. The two raters then classified all instances independently. They proposed the same classification for 201 instances (91%). They considered the 21 cases of disagreement in more detail, in some cases taking a wider context around the repair into consideration, and reached a consensus classification for 15. The remaining 6 instances, for which the raters agreed that either classification could be proposed, were excluded from further analysis — which leaves the 216 instances on which we report in this paper. Among these 216 instances, error repairs outnumber appropriateness repairs (N=129 and N=87, respectively).

In order to assess whether factual and linguistic errors give rise to different repair prosodies, given the distinct levels of processing involved in error detection, the first author further classified the 129 confirmed error repairs accordingly. It was deemed unnecessary to involve the second rater in the further classification, as this could be partly based on notes recorded by both raters for the purpose of the main classification. (We will refer to this variable as *Error type*.) All instances in which the reparandum item would have resulted in a clearly ill-formed collocation, as in (1b) were classified as linguistic errors; all others,

including (1a), (1d) and (1g), as factual error repairs. Among the 129 error repairs, factual error repairs outnumber linguistic error repairs (N=93 and N=36, respectively).

Semantic field size and frequency

Assessing the role of semantic field size in conditioning repair prosody is less straightforward, as establishing numbers of contextual alternatives to the reparandum item is in most cases impossible. However, we could identify a subset of 29 error repairs with a clear maximum semantic field size. These include repairs involving antonyms, in which the reparandum item can be said to operate in a semantic field comprising just two items; days of the week, in which the field comprises seven items; up to a maximum field size of 12 for months of the year.

In addition, we took several frequency measurements, on the assumptions that frequent items are more predictable than infrequent ones, and more predictable items can be seen as items with fewer contextual alternatives than less predictable items — and given the expected interaction between word frequency and repair timing described above. We took word and lemma frequency counts for the reparandum and repair items from CELEX [Baayen et al. 1995]. (We will refer to these variables as *Reparandum lemma frequency*, *Repair word frequency*, and so on.) In addition to entering the (log-transformed) counts straight into our quantitative analysis, we subtracted the reparandum count from the repair count to yield a measure of the frequency differential between the two lexical items involved in the repair. (We will refer to these variables as *Lemma frequency delta* and *Word frequency delta*.) Positive values correspond to a repair item that is more frequent than the item it replaces; negative values to a repair item that is less frequent.

Repair timing

In order to assess whether repairs with a reparandum item that is interrupted early have different prosodic characteristics from repairs with a completed reparandum item, following Nooteboom's [2010] findings on phonological error repairs, we classified each reparandum item as interrupted or completed prior to repair, as illustrated in (1). (We will refer to this variable as *Completeness*.) All morphologically complex words, including compounds, were treated as single words for this purpose: in other words, (1g) is considered interrupted even though the crucial reparandum morpheme, *computer*, is a free morpheme and is completed prior to the repair. Such complex reparandum items constitute less than 10% of the data set,

and exploratory analysis (not reported here) suggested that treating them differently would not alter the main findings reported below.

In addition to a binary measure of repair timing, we explored the relevance of two continuous measures, on the assumption that these might capture more fine-grained differences between 'early' and 'late' repairs. First, we measured the duration from the start of the reparandum item to the abandonment of speech prior to repair: all other things being equal, the longer this interval, the later the repair. (We will refer to this variable as *Reparandum duration*.) Second, we took a proportional measure of reparandum item completeness. This is appropriate since our reparandum items are not independently controlled for word length (unlike in Nooteboom's 2010 study) or speaking rate. As a result, a duration measurement only partially captures repair timing: it may be that what matters most is how much of the reparandum item has been completed prior to repair, irrespective of how long it has taken the speaker to do this.

To implement the proportional measure, we divided the number of segments produced between the start of the reparandum item to the abandonment of speech prior to repair by the number of segments in the (projected or completed) reparandum item. (We will refer to this variable as *Proportional completeness*.) We ignored segment deletions for this purpose: the crucial question was which segment in a canonical realisation of the word in question was reached in the surface form. We referred to Heemskerk and Zonneveld [2000] for the segmental make-up of all canonical forms. Note that the measure is not bounded by 1: instances in which the speaker produces further lexical material following the reparandum item, but prior to repair, result in values above 1. All other things being equal, the higher the value, the later the repair.

Statistical modelling

In what follows, we will first examine the relationship between the various f0 and intensity measures and prosodic marking judgements outlined above, and then present results of attempts to establish the predictive value of our semantic, temporal and frequency-related factors. We mainly present results of analyses using conditional inference regression trees. Given a dependent, or 'response' variable and a set of candidate predictor variables, the algorithm establishes which predictor variables give rise to homogeneous sub-groupings of observations with respect to the levels of the response variable, and outputs a tree diagram in which each predictor variable that does give rise to a sub-grouping is represented as a node. The algorithm works recursively, in that given the identification of multiple significant

predictors in a data set, the data is first split into two subsets according to the strongest predictor. Each of the resulting subsets is then inspected to establish whether other predictors give rise to further, subordinate groupings. This way, predictor interactions emerge as asymmetrically nested nodes; we will see one example of this below.

An independent variable that does not give rise to any 'splits' in the data can be compared to a non-significant factor in a linear regression model. Moreover, for each tree, a coefficient such as C or r^2 can be computed as an indicator of the proportion of variance in the data that the model accounts for, as in the case of linear models. However, as pointed out by Strobl et al. [2009] and Tagliamonte and Baayen [2012], analysis based on conditional inference trees has the important advantage over linear regression modelling that it is highly robust in the face of collinearity among predictors, which can give rise to spurious main effects and interactions in linear models — or requires elaborate stepwise modelling procedures to avoid such spurious results. Since many of our candidate predictors are expected to be highly correlated with each other — for example, because they are alternative measures of the same basic parameter, such as repair timing or lexical frequency — analysis based on conditional inference trees is a useful alternative to linear modelling. (We did construct linear mixed-effects models for most tree-based models we report on below, and these do not give us reason to doubt the robustness of the tree-based models. In some cases they point towards complex interactions between, or even contradictory main effects of highly correlated predictors which are not reflected in the tree-based models. We assume these are uninformative effects of collinearity.)

RESULTS

Turning now to the results of our analyses, we first examine the relationship between the various f0 and intensity measures we took and the prosodic marking judgements, and the relationship between our semantic, temporal and frequency-related predictor variables. We then evaluate the predictive value of the predictor variables in modelling the prosodic variables, and, as a control procedure, also use the prosodic variables to model some of our main predictors. Finally, we report on our attempt to model semantic field size in the small subset of data for which this can be quantified.

Acoustic measures and prosodic marking judgements

As indicated above, our auditory analysis resulted in 43 instances (20%) being classified as prosodically marked, 24 (11%) as possibly marked and 149 (69%) as prosodically unmarked.

In order to assess the relationship between this classification and our acoustic measures of f0 and intensity maximum, median and mean, we can first consider the distributions of the six delta measures, given in Figure 2. If these accurately reflect that prosodic marking is rare in our data, they should show clear peaks centred around 0, reflecting an absence of change between the reparandum and repair items on the parameter in question. Moreover, if in the subset of marked instances, marking is achieved either by a considerable increase in pitch or intensity or a considerable decrease, as suggested by Cutler [1983], the distributions of individual parameters may show evidence of multimodality. On the other hand, if some degree of f0 and intensity raising is the norm, as found by Howell and Young [1991] and Nakatani and Hirschberg [1994], the distributions are most likely to be unimodal and show evidence of negative skew.

Figure 2 shows that indeed, most distributions have a clear peak around 0. The peaks are very sharply defined in the case of the f0 measures, which means that in the majority of instances, the pitch characteristics of the repair are close to identical to those of the reparandum. The peaks have broader bandwidths for the intensity measures, suggesting wider spreads of values, and there is a hint of a 'right shoulder' in all three distributions. This means a fairly sizeable subset of instances involve a moderate rise in intensity. This may be taken as weak evidence of negative skew. On the other hand, there is no clear evidence of multimodality in our delta measures, as confirmed by Hartigans' dip tests (across parameters, D ranges between 0.0118 and 0.0204, $p>0.8$).

Figure 3 shows corresponding f0 and intensity delta measures plotted against each other, with prosodically marked, possibly marked and unmarked instances labelled separately. If f0 and intensity are manipulated independently in the prosody of self-repair, as suggested by Cutler's [1983] definition of prosodic marking, we would expect data points to fall into more or less discrete clouds. Moreover, if our chosen acoustic parameters are among those on which the perception of prosodic marking is based, we would expect data points representing marked, possibly marked and unmarked instances to cover distinct subareas of the plots. Concretely, Cutler's [1983] definition of prosodic marking suggests we should expect marked instances to cluster around the periphery of the plots, where data points represent instances with a large absolute delta value for one or both acoustic parameters.

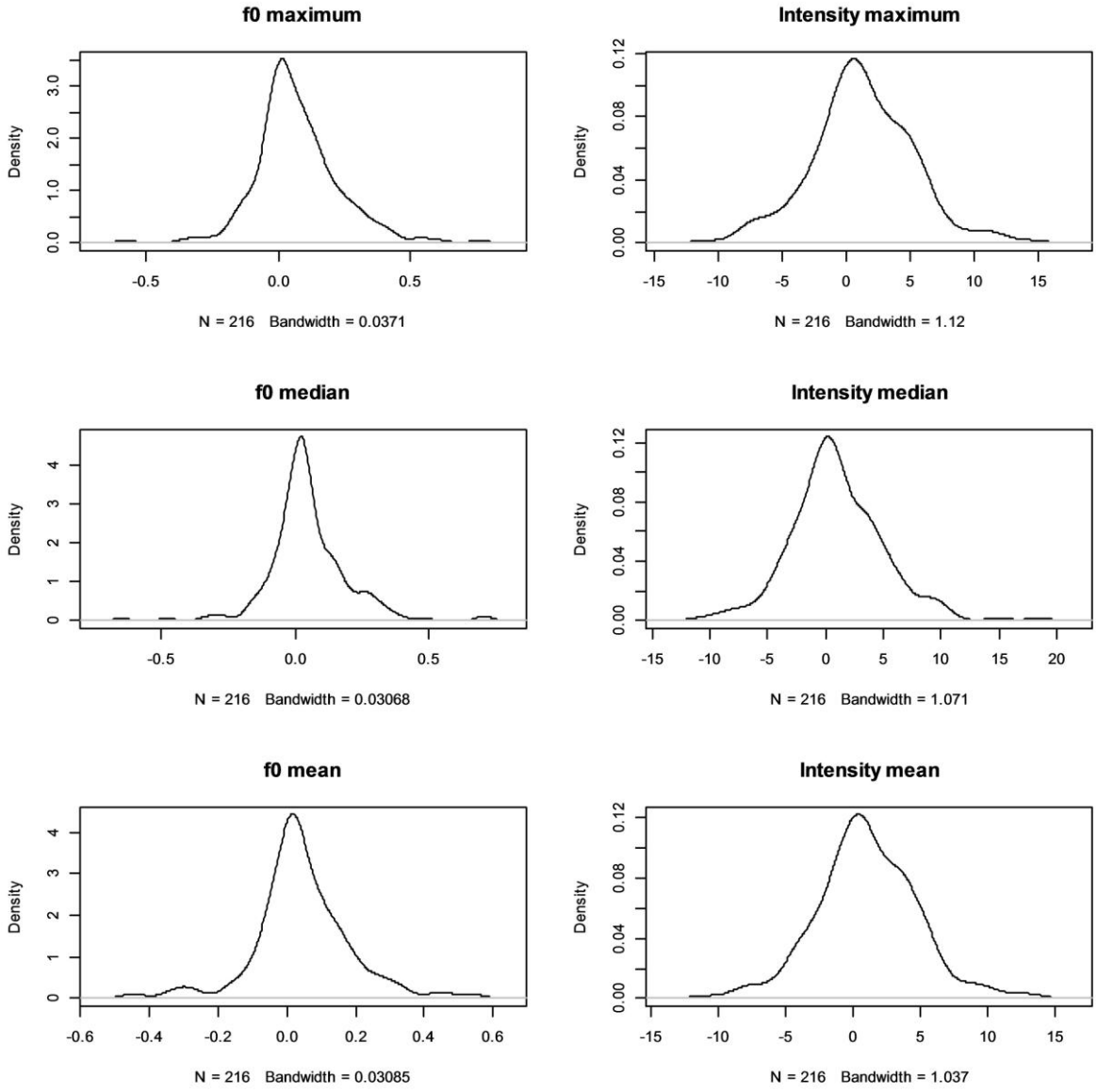
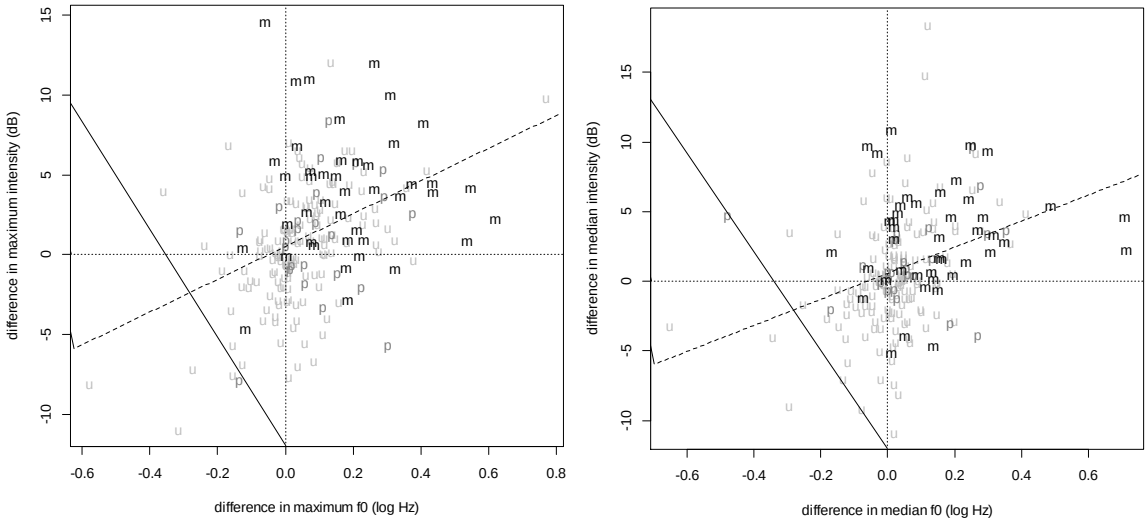
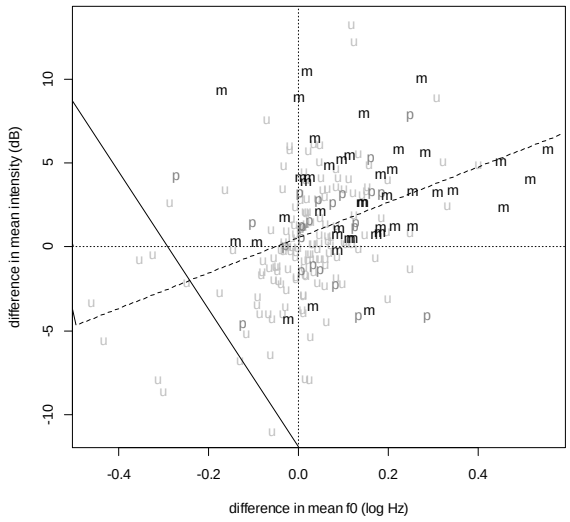


Figure 2. Kernel density plots for f0 maximum, median and mean deltas (left) and intensity maximum, median and mean deltas (right).



(a) (b)



(c)

Figure 3. Scatterplots of intensity deltas by f0 deltas for maxima (a), medians (b) and means (c), split by *Prosodic marking*. Data points plotted with a black *m* represent instances classified as 'marked'; data points plotted with a dark grey *p* represent instances classified as 'possible marked'; and data points plotted with a light grey *u* represent instances classified as 'unmarked'. Slopes drawn in a dotted line represent the outcome of a simple linear regression model in each case.

Looking first at the relationship between our f0 and intensity measures, as expected on the basis of the distributions shown in Figure 2, for each of the three parameters the majority of instances have delta values around 0 for both f0 and intensity. Moreover, most of the scatters show what looks like a single cloud of data points with a positive correlation between the two dimensions (maximum: $\rho=0.4365$, $p<0.0001$; median: $\rho=0.3928$, $p<0.0001$; mean: $\rho=0.4338$, $p<0.0001$; we use Spearman's rho since none of the distributions is normal, as confirmed by Shapiro-Wilks tests). These correlations mean that a repair produced with a rise in f0 mostly has a rise in intensity too; conversely, a repair produced with a fall in f0 mostly has a fall in intensity.

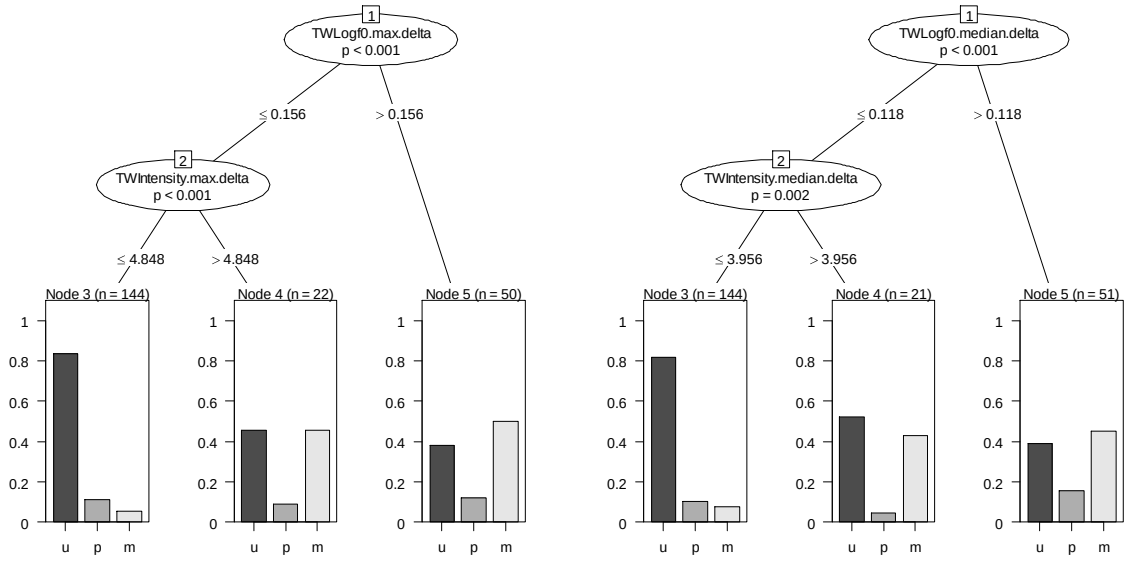
With respect to the relationship between our acoustic measures and the auditory judgements, Figure 3 shows that for each of the three parameters, the vast majority of data points corresponding to instances that are perceived as prosodically marked occupy the top right quarter of the plot. (Of instances classified as 'marked', between 86% and 91% have a positive delta value, depending on the acoustic parameter. Of instances classified as either 'possibly marked' or 'marked', between 78% and 87% have a positive delta value, depending on the acoustic parameter.) These data points represent instances with a rise in f0 and intensity between the reparandum item and the repair item. Instances with a fall in f0 and intensity do occur in our data set, as seen in the bottom left quarters of the plots; however, very few of these were perceived by our raters as (possibly) marked.

While most instances that are perceived as prosodically marked have positive delta values for f0 and intensity, it is not clearly the case that the majority of instances with positive delta values for f0 and intensity are perceived as prosodically marked. Of all instances with positive delta values, only between 25% and 29% were classified as 'marked', depending on the acoustic parameter. Between 37% and 42% were classified as either 'marked' or 'possibly marked', depending on the acoustic parameter. In other words, on each of the six acoustic parameters, over half of the instances with a positive delta value are perceived as prosodically unmarked. It is also not the case that the data points occupy particularly peripheral subareas of the plots, the distributions do suggest that the higher the increase in f0 and intensity maximum, median and mean between a reparandum and repair item, the greater the likelihood that the repair is perceived as prosodically marked.

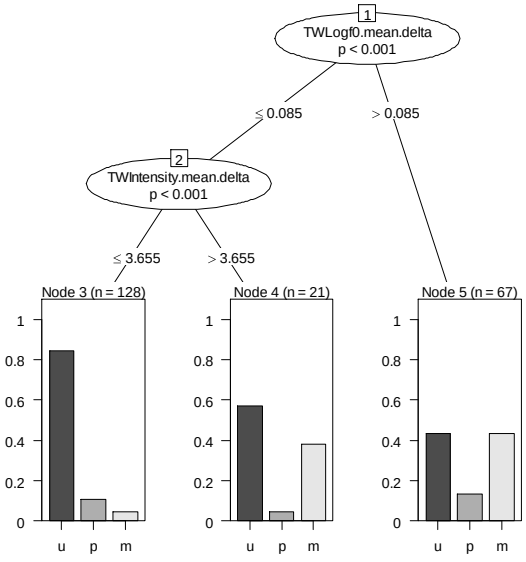
These observations are confirmed by further statistical analysis. Figure 4 shows three conditional inference regression trees, each modelling the prosodic marking judgements (coded as an ordinal factor with three levels: *unmarked*, *possibly marked*, then *marked*) on the basis of two corresponding f0 and intensity delta measures. The trees are very similar,

showing a first split of the data on the f0 variable in question, and a second split on the intensity variable, reflecting three homogeneous subsets of data. The first contains between 128 and 144 instances with pitch delta values up to about 0.1 and intensity delta values up to about 4. The second is a small subset (21 or 22 instances depending on the tree) with pitch delta values up to about 0.1 and intensity delta values above about 4. The third has pitch delta values above about 0.1, and contains between 50 and 67 instances. The bar charts at the bottom of the trees show that of the instances in the first subset (left), with relatively low pitch and intensity deltas, about 80% are classified as prosodically marked. Of the instances in the second subset (middle), with relatively low pitch deltas and relatively high intensity deltas, either a small majority are classified as unmarked (median, mean) or equal proportions are classified as marked and unmarked (maximum). Of the instances in the third subset (right), with pitch maximum deltas above about 0.1, either a small majority are classified as marked (maximum, median) or equal proportions are perceived as marked and unmarked (mean).

The trees in Figure 4 confirm that there is a significant relationship between our acoustic parameters and prosodic marking judgements, such that the higher the increase in f0 and intensity maximum, median and mean between a reparandum and repair item, the greater the likelihood that the repair is perceived as prosodically marked. The trees allow for between 69% and 72% of the data to be correctly classified with respect to *Prosodic marking* on the basis of the acoustic measurements, with the tree for maximum delta values performing best (71.8%). Subsequent modelling using random forests [Breiman 2001], which allows for the calculation of relative importance among correlated predictor variables [see Tagliamonte and Baayen 2012], suggests that *f0 mean delta*, *f0 maximum delta* and *Intensity maximum delta* are the strongest predictors of our prosodic marking judgements, followed at some distance by *f0 median delta*. This modelling also suggests that *Intensity median delta* and *Intensity mean delta* do not constitute significant predictors on their own.



(a) (b)



(c)

Figure 4. Conditional inference regression trees predicting prosodic marking judgements on the basis of f0 and intensity maximum delta values (a), median delta values (b) and mean delta values (c).

Relationships among predictor variables

Prior to modelling the acoustic measures and prosodic marking judgements using the semantic, temporal and frequency-related factors described above, we assessed the extent to which these factors are correlated. As indicated above, some correlations have been noted in previous literature: Levelt [1989: 481] and Brédart [1991] report that error repairs are more likely than appropriateness repairs to involve an early interruption of the reparandum item, while Kapatsinski [2010] report that high-frequency reparandum items are less likely than low-frequency items to involve an early interruption.

Our data provide no support for Levelt and Brédart's findings: all cross-tabulation and regression models involving either *Completeness*, *Proportional completeness* or *Reparandum duration* on the one hand and *Repair type* or *Error type* on the other produce non-significant results (for example, for *Completeness* and *Error type*, χ^2=1.8502, df=2, p=0.3965; for *Reparandum duration* and *Error type*, a linear regression model yields R^2=0.0032, p=0.7080). This means that in our data, error and appropriateness repairs show no difference in the likelihood of the reparandum item being interrupted prior to repair, and the relevant factor groups can be considered fully independent for the purpose of modelling repair acoustics and prosodic marking judgements.

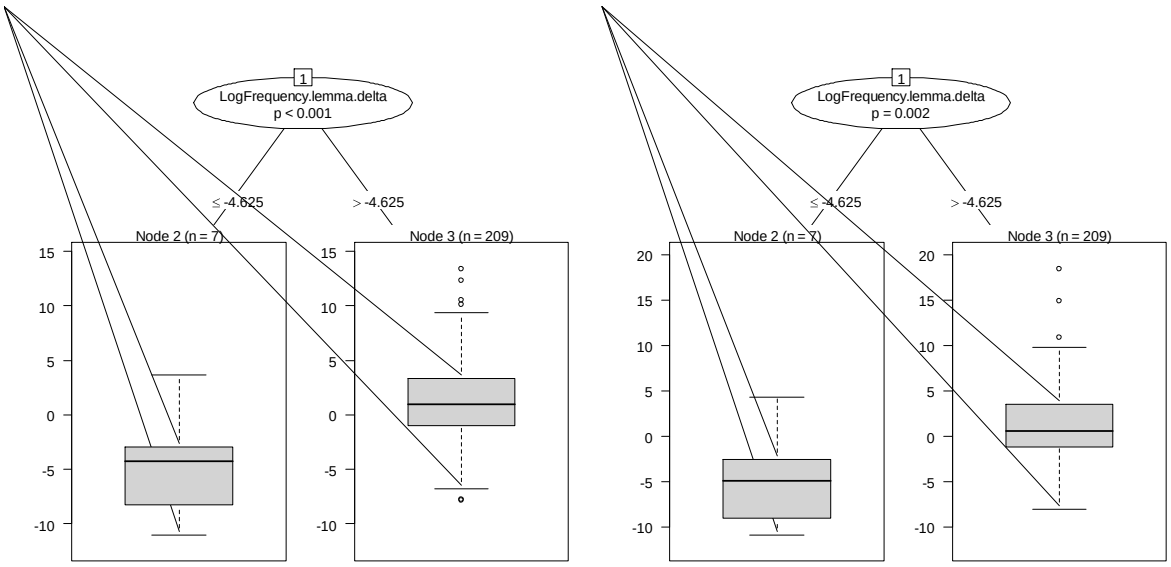
Our data do provide support for Kapatsinski's finding of a negative relationship between lexical frequency and the likelihood of an early interruption in repair: for example, we find significant correlations between *Reparandum duration* and *Word frequency* (Spearman's ρ=-0.1762, p=0.0094) as well as *Lemma frequency* (Spearman's ρ=-0.1885, p=0.0054). This means that relevant factor groups cannot be considered fully independent for the purpose of modelling repair acoustics and prosodic marking judgements.

Modelling the acoustic measures and marking judgements

Turning now to the predictive value of the semantic, temporal and frequency-related factors described above, we first attempted to model each of the six acoustic parameters and *Prosodic marking* using conditional inference regression trees. In each case, we entered the candidate predictors *Repair type*, *Error type*, *Completeness*, *Proportional completeness*, *Reparandum duration*, *Reparandum lemma frequency*, *Repair lemma frequency*, *Lemma frequency delta*, *Reparandum word frequency*, *Repair word frequency*, and *Word frequency delta*. In addition, we included four control variables to take account of any effects of speaker identity and language variety and style: first, the speaker's name; second, the speaker's

gender; third, the subcorpus from which each instance of repair was extracted; and fourth, the variety of Dutch spoken (Netherlands Dutch versus Flemish).

Our analysis revealed very few significant effects of our candidate predictors. There were no splits in the tree for *Prosodic marking*, and no splits in any of the trees based on f0 measures, suggesting that none of our semantic, temporal, frequency-related or control variables have any significant effect on the changes in f0 between reparandum and repair or on the likelihood of an instance of repair being perceived as prosodically marked. Two of the three intensity measures reveal one significant split each in the data. Further inspection suggests that we are looking at one significant effect of a frequency-related variable: as shown in Figure 5, *Lemma frequency delta* yields identical homogeneous sub-groupings for intensity median and mean deltas. The effect is weak: Figure 5 shows that it consists in a subset of 7 instances with a particularly low negative value for *Lemma frequency delta* — that is, a particularly large decrease in lemma frequency from reparandum item to repair item — having a significantly lower drop in intensity median and mean than the rest of the data set. Values for r^2 for the tree predictions are 0.07 for the intensity median deltas and 0.08 for the intensity mean deltas, which suggests the effect accounts for at most 8% of the variance in the data.



(a) (b)

Figure 5. Conditional inference regression trees predicting the difference in intensity mean (a) and intensity median (b) between reparandum and repair items. In both cases, the data are split on the variable *Lemma frequency delta*.

Given its weakness, we cannot draw firm conclusions from the observed frequency effect. Arguably more pertinent is the finding that neither the error–appropriateness dichotomy, nor the distinction between 'early' and 'late' repairs appears to have any significant effect on the f0 and intensity contours associated with the repairs, or on the likelihood of the repair being produced with noticeable prosodic marking. The finding that repair timing does not have a significant effect on the likelihood of prosodic marking is in line with both Levelt and Cuter's [1983] and Cutler's [1983] results; however, the finding that it has no significant effect on f0 and intensity deltas goes against Nooteboom's [2010]. The finding that neither *Repair type* nor *Error type* has a significant effect on the likelihood of prosodic marking goes against Levelt and Cutler's [1983] results.

In order to ensure that our negative finding regarding the relationship between repair semantics and prosody is not due only to our use of different statistical methods from those of Levelt and Cutler [1983], we also replicated their method, which involved the use of simple cross-tabulation tests only. The results are shown in Table 1, which confirms that *Repair type* has no significant effect on the likelihood of prosodic marking, whether our 'possibly marked' classification is treated as a separate level, or collapsed with 'marked' or 'unmarked'. However, *Error type* does appear to have a significant effect, both when 'possibly marked' is treated as a separate level and when it is collapsed with 'marked'. Figure 6 suggests that the significance is due to a comparatively high likelihood for repairs of factual errors to be prosodically marked — in particular when our classifications 'marked' and 'possibly marked'are both taken to reflect a degree of prosodic marking. Repairs of linguistic errors, on the other hand, show very similar proportions of 'marked', 'possibly marked' and 'unmarked' instances to appropriateness repairs. The fact that the effect does not yield a corresponding split in the conditional inference regression tree for *Prosodic marking* suggests the effect is again a weak one.

Moreover, in order to ensure that the great number of instances in which there is very little prosodic change between the reparandum and repair items do not mask effects of our semantic and temporal variables — or in other words, to establish whether any effects can be observed at the peripheries of the data scatters in Figure 3, we reconstructed the conditional inference tree for each prosodic parameter three times: once removing instances that are less than 1 standard deviation away from the mean delta value, once removing instances that are less than 1.5 standard deviations away, and once removing instances that are less than 2 standard deviations away. None of the resulting trees reveal any significant splits. We attempted a variety of additional measures to focus on peripheral instances, including

measures which collapsed f0 and intensity measures together, such as Euclidean distance from zero and principal components analysis. None of these methods revealed any additional significant effects.

	Prosodic marking levels	χ^2	df	p
(a)	M *vs* P *vs* U	3.1907	2	0.2028
	M *vs* P, U	0.0811	1	0.7759
	M, P *vs* U	1.8102	1	0.1785
(b)	M *vs* P *vs* U	10.9854	4	0.0267
	M *vs* P, U	0.7795	2	0.6772
	M, P *vs* U	7.7992	2	0.0203

Table 1. Results of Pearson's chi-squared tests, with Yates' continuity correction where appropriate, for (a) *Repair type ~ Prosodic marking* and (b) *Error type ~ Prosodic marking*. Under '*Prosodic marking* levels', 'M' stands for 'marked', 'P' for 'possibly marked', and 'U' for 'unmarked'. Each test was run with 'possibly marked' as a separate level, collapsed with 'unmarked' and collapsed with 'marked', respectively.

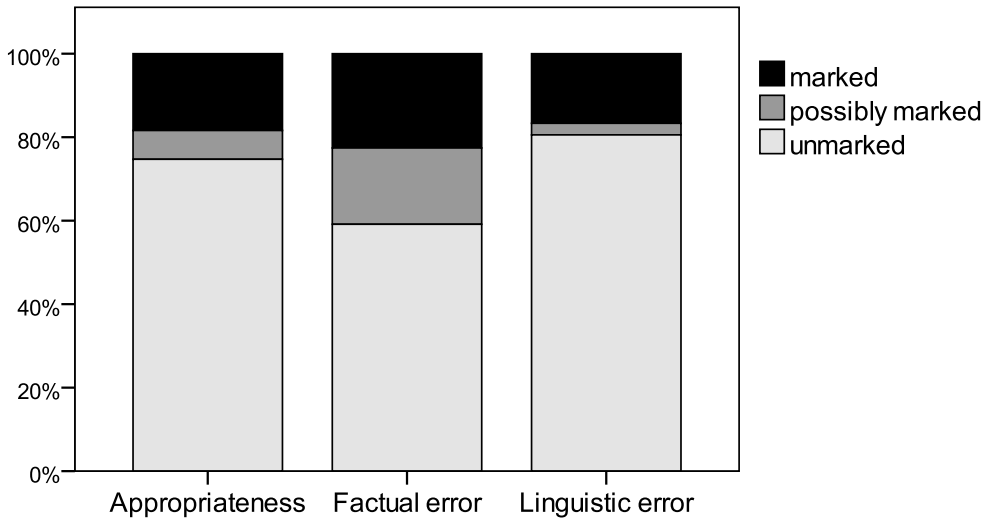


Figure 6. Descriptive statistics for *Error type ~ Prosodic marking*: relative frequencies of marked, possibly marked and unmarked instances for appropriateness repairs, factual error repairs and linguistic error repairs.

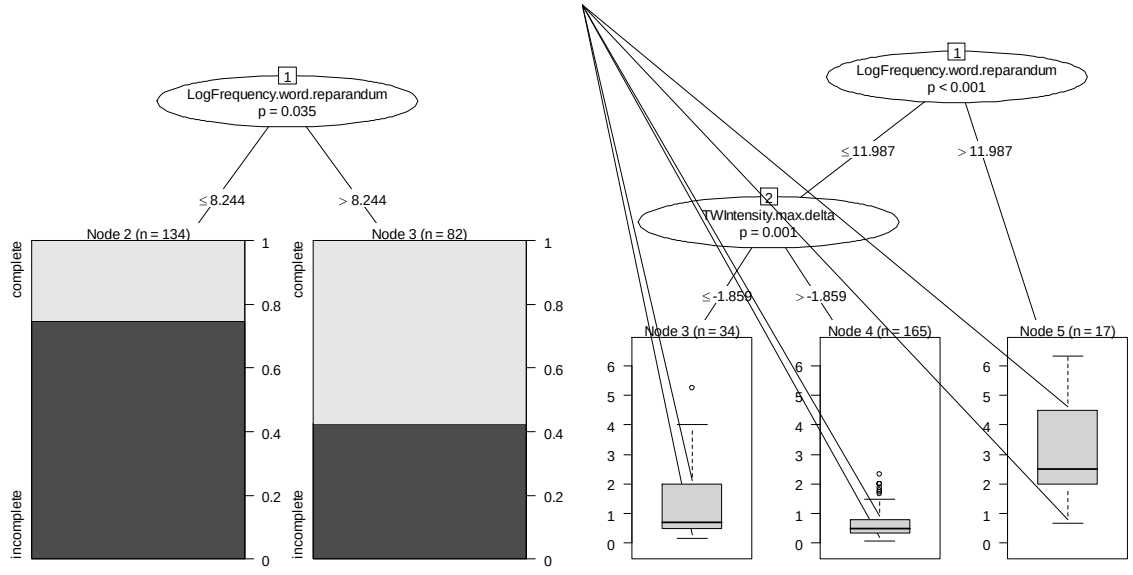
Modelling the semantic and temporal variables

In order to make sure that our negative findings with respect to the predictive value of semantic and temporal variables are robust, and to rule out the possibility that our modelling prosodic parameters separately masks subtle effects of semantic or temporal variables across parameters, we constructed four further conditional inference trees. The first had *Repair type* as response variable, and all of our six acoustic parameters and *Prosodic marking* as candidate predictor variables. We also included the temporal and frequency-related variables described above among the candidate predictors. The second, third and fourth trees were construed along similar lines for *Completeness*, *Proportional completeness* and *Reparandum duration*: again, the prosodic parameters were our crucial candidate predictors, and in these cases we included semantic and frequency-related variables to capture any significant relationships among our original predictor variables.

The analysis confirms our previous negative findings with respect to the relationship between prosodic parameters on the one hand and semantic, temporal and frequency-related parameters on the other. Of the four trees, only one — the one for *Proportional completeness* — contains a split on a prosodic parameter. We will return to this below. With respect to the relationship between semantic, temporal and frequency-related parameters, the analysis confirms the findings reported so far. That is, none of our variables provide a handle on the error–appropriateness dichotomy: the conditional inference tree for *Repair type* contains no significant splits. The same appears to be the case for *Reparandum duration*, but *Completeness* and *Proportional completeness* can be predicted to some extent using a combination of prosodic and frequency-related predictor variables. The conditional inference trees are given in Figure 7.

Figure 7 shows that in modelling binary *Completeness*, the data can be subdivided according to *Reparandum word frequency*, such that in a subset of 82 instances with a high reparandum word frequency, the proportion of completed reparandum items is significantly higher than in the rest of the data. This means that a high reparandum word frequency increases the likelihood of the reparandum item being completed prior to repair, as also reported by Kapatsinski [2010]. The same split emerges in modelling *Proportional completeness*, although in this case the subset of high-frequency reparandum items — associated with significantly greater proportional completeness values than the rest of the data — is much smaller, at 17. Interestingly, in modelling *Proportional completeness* a further split emerges in the rest of the data. That is, if we disregard the 17 instances with a particularly high-frequency reparandum item, it appears that a subset of 34 instances with a

particularly low negative intensity maximum delta — in other words, a large drop in intensity maximum from reparandum item to repair item — are associated with higher proportional segment counts for the reparandum item than the rest of the data.



(a) (b)

Figure 7. Conditional inference regression trees predicting binary *Completeness* (a) and *Proportional completeness* (b). In both cases, the data are split on the variable *Reparandum word frequency*; in the case of *Proportional completeness*, a further split is possible on *Intensity maximum delta*.

As in the case of the models in Figure 5, the significant splits in the models in Figure 7 separate relatively small subsets of instances from the rest of the data set. Unsurprisingly, the overall model prediction is unimpressive in both cases: the index of concordance, C, for the *Completeness* model is 0.6605, indicating marginally better than chance prediction. With its interaction between *Reparandum word frequency* and *Intensity maximum delta*, the *Proportional completeness* model is the most comprehensive model generated in our analyses. Still, its r^2 of 0.4325 indicates that not even half of the variance in the data is accounted for. Again, then, it is difficult to draw firm conclusions about the effects included in the models. We will return to their interpretation below.

Modelling semantic field size

Finally, we considered the subset of 29 instances for which we can establish a semantic field size, as explained above. As before, we constructed conditional inference regression trees with each of our acoustic parameters and *Prosodic marking* as response variable. This time, we only included the (log-transformed) field size as a candidate predictor variable. There were no significant splits in any of the trees, which means we have no clear evidence of a link between the size of the semantic field from which the reparandum and repair are drawn and the change in f0 and intensity between reparandum and repair, or the likelihood of the repair being prosodically marked. We also built simple linear regression models with the (log-transformed) field size as the only predictor variable, and mixed models with the additional random factor *Speaker*. None of these revealed significant effects. Still, visual inspection of the distributions in question is suggestive of a predictable relationship between field size and likelihood of prosodic marking, which might not emerge as significant because of the small size of the data set in our study. As shown in Figure 8, instances classified as 'marked' have a lower median field size than instances classified as 'possibly marked', which in turn have a lower median field size than instances classified as 'unmarked'. The direction of this tendency is consistent with that reported by Levelt and Cutler [1983]: the smaller the field size, the greater the likelihood of prosodic marking.

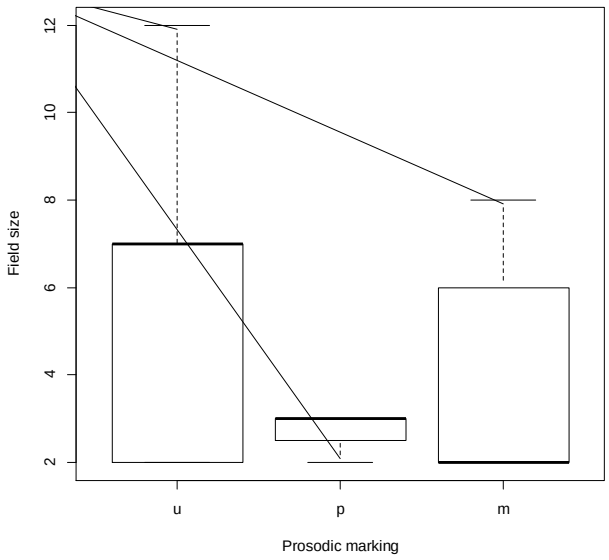


Figure 8. Boxplot illustrating the relationship between *Prosodic marking* and *Field size*. For *Prosodic marking*, 'M' stands for 'marked', 'P' for 'possibly marked', and 'U' for 'unmarked'.

DISCUSSION

In this paper we have reported on a phonetic analysis of instances of lexical self-repair, focusing on how the pitch and intensity of the second — preferred — lexical item compare to those of the first — rejected — one, and on the relevance of semantic, temporal and frequency-related variables in modelling this relationship. In this section, we first discuss our findings with regards to the overall frequency of prosodic marking and the relationship between auditory judgements and our acoustic measurements, and then turn to the relevance of our various candidate predictors.

Prosodic marking, pitch and intensity

A first discussion point is the low observed proportion of prosodically marked instances, compared with the proportions reported by Cutler [1983] and, in particular, Levelt and Cutler [1983]. When only those instances classified as 'prosodically marked' are considered, the observed proportion (20%) is 18% below that reported by Cutler and 25% below that of Levelt and Cutler. When instances classified as 'prosodically marked' and 'possibly marked' are binned for comparison, the resulting proportion (31%) is still 7% below Cutler's and 14% below Levelt and Cutler's. It is of course possible that the auditory analysis we conducted was more conservative than Levelt and Cutler's. Neither Cutler [1983] nor Levelt and Cutler [1983] provide a detailed description of their auditory analysis procedure, so it is unclear how successfully we have replicated it. However, our acoustic analysis results are consistent with the low observed proportion of prosodically marked instances, in that a majority of instances involve very little change in either f0 or intensity between the reparandum item and the repair item. Moreover, it is notable that our observed proportions are closer to those of Cutler [1983] than to those of Levelt and Cutler [1983]: like our instances of self-repair, Cutler's were drawn from spontaneous speech material, as opposed to the task-oriented dialogue of Levelt and Cutler [1983]. It seems plausible that speakers' responses to issues of factual accuracy, linguistic well-formedness and pragmatic felicity might be different in an explicitly task-oriented setting as compared with an unconstrained spontaneous speech setting.

In addition, our analysis shows that the proportion of 'marked' instances is to some extent constrained by the relative proportions of factual error, linguistic error and appropriateness repairs; see below for further discussion. If our observation that factual error repairs are more frequently prosodically marked than linguistic error and appropriateness repairs proves generalisable to other data sets, differences in the prevalence of prosodic marking in a given collection of lexical repairs may be at least partly attributable to

difference in the relative proportions of the three subtypes of repair. Unfortunately, Levelt and Cutler [1983] and Cutler [1983] do not distinguish between factual and linguistic error repairs, but intuitively it does not seem implausible for Levelt and Cutler's network description task to have elicited a high proportion of errors of 'fact' — colour, direction, shape and so on — relative to linguistic formulation errors. Similarly, if semantic field size proves a consistent, if weak, predictor in further work, the relative proportion of repairs involving antonyms or highly restricted semantic domains can be expected to have an impact on overall rates of prosodic marking across data sets.

With reference to our acoustic measurements, we saw some evidence of negative skew, in particular in the intensity distributions: on the whole, a majority of instances involve a moderate rise in intensity between the reparandum and repair items, as also reported by Howell and Young [1991] and Nakatani and Hirschberg [1994]. We saw no evidence of multimodality in any of the distributions, and mapping the auditory marking judgements to the acoustic measurements showed that the majority of prosodically marked instances involve a rise in f0 and intensity between reparandum and repair items. In other words, like Hokkanen [2001] and Cole et al. [2005], we find little evidence that prosodic marking in self-repair is achieved through a noticeable fall along either f0 or intensity. This is consistent with Cutler's [1983: 80–81] assertion that 'typically', a marked repair 'is uttered on a higher pitch and with greater intensity than the erroneous material', and suggests that her reference to repairs being marked 'by being uttered on a noticeably lower pitch' is relevant to a small minority of instances only, if any.

We also saw that repairs that *are* uttered on a lower pitch than the reparandum tend to be uttered on a lower intensity too: the two acoustic parameters of f0 delta and intensity delta show a significant positive correlation, and analysis using conditional inference regression trees confirms that each is a major predictor of the other. While we have not investigated tempo in this paper, we can conclude that the independence of pitch and intensity parameters implied by Cutler's and Levelt and Cutler's definitions of prosodic marking in repair — as involving 'a noticeable increase or decrease in pitch, in amplitude, or in relative duration' [Levelt and Cutler 1983: 206] or 'longer relative duration, noticeably higher or lower pitch, noticeably higher or lower amplitude, or a combination of pitch, amplitude and durational effects' [Cutler 1983: 84] — should not be overestimated. In our data, pitch and intensity are by and large manipulated in tandem, not independently. This is consistent with Nooteboom's [2010] findings on the phonetic differentiation of speech error repairs, as well as with those of various studies of sound patterns in spontaneous interaction.

As suggested above, researchers combining phonetic analysis and conversation-analytic methods tend to emphasize the importance of detailed analysis of the 'clusters' of phonetic features that give rise to auditory impressions of 'emphasis', 'foregrounding', 'prosodic marking' and so on, guided by the notion that there is no *a priori* way of predicting how these clusters will be constituted in any given interactional context [Local 2003, Local and Walker 2005, Selting 2010: 27]. Indeed, we know that speakers *can* manipulate f0 and intensity independently, as shown for example by comparisons between infant-directed, Lombard and 'clear' speech [Wassink et al. 2007, Smiljanić and Bradlow 2009]. Still, studies of sound patterns in interaction have repeatedly found associations between high pitch and high intensity on the one hand, and low pitch and low intensity on the other: see Walker [2009], Ogden [2006, 2010] and Local et al. [2010] for recent examples, attested in a range of communicative contexts. Our findings add that of prosodic marking in self-repair.

Returning to the relationship between auditory prosodic marking judgements and our acoustic measures of f0 and intensity deltas, as pointed out above, the majority of prosodically marked instances involve a rise in f0 and intensity between reparandum and repair items. However, as pointed out above, it is *not* the case that a clear majority of instances with positive delta values for f0 and intensity are perceived as prosodically marked, and analysis using conditional inference regression trees reveals that the three acoustic parameters account for at most 72% of the auditory prosodic marking judgements. This suggests that while f0 and intensity maximum, median and mean are useful parameters for capturing the auditory judgements, they do not capture them entirely. Tempo, voice quality and articulatory setting are among additional parameters that may be relevant [see Niebuhr 2010], and it may be that alternative measures of in particular intensity, such as spectral tilt or root-mean-square amplitude [see e.g. Sluijter and Van Heuven 1994] produce a better fit to the auditory judgements. We are also aware that our reliance on two raters only and method of dealing with initially non-matching judgements may have introduced noise in our data. Further research is needed to address these issues.

Predictive value of semantic, temporal and frequency-related variables

As indicated at the outset of this paper, Levelt and Cutler [1983] report a significant difference in the frequency of prosodic marking between error and appropriateness repairs, as well as a significant effect of semantic field size. Neither effect is found to be significant in our data, although tendencies that fit with Levelt and Cutler's findings can be observed. First, the data show a weak tendency for repairs of factual errors to be prosodically marked more

frequently than repairs for appropriateness reasons. This tendency is only visible when factual error repairs are distinguished from linguistic error repairs, which are marked as frequently as appropriateness repairs. This raises the possibility that whether or not the error–appropriateness distinction is a significant factor in modelling repair prosody depends to some extent on the relative proportions of factual and linguistic error repairs: if the proportion of factual error repairs is high enough, a difference between these repairs on the one hand and appropriateness and linguistic error repairs on the other may surface as a significant effect of the error–appropriateness distinction, masking the difference among the two subtypes of error repair. Of course, further work is needed to establish whether the observed tendency generalises beyond our data; the fact that it does not yield a significant effect in a conditional inference regression model suggests this may not be the case. As indicated above, unfortunately Levelt and Cutler [1983] do not specify the relative proportions of factual and linguistic error repairs.

Second, while semantic field size does not yield significant effects in our data, the data set is very small, and the descriptive statistics are as Levelt and Cutler [1983] would predict: 'marked' instances have a lower median semantic field size than 'possibly marked' instances, which in turn have a lower median than 'unmarked' instances. This is consistent with the idea that the smaller the number of lexical competitors, the more 'contrastive' the repair is, and therefore the greater the likelihood of prosodic marking to foreground the correct lexical choice. Again, further research is needed is this area. One possibility is to elicit self-repairs in an experiment similar to that of Hartsuiker and Notebaert [2010]. Hartsuiker and Notebaert use a picture-naming task to investigate whether 'name agreement' — the number of alternative names for a given object [see Severens et al. 2005] — is a significant predictor of the likelihood of disfluency in naming the corresponding picture. They find that it is, but do not consider the phonetic characteristics of the elicited disfluencies in any detail. Based on Levelt and Cutler [1983], we might predict that prosodic marking is most common among self-corrections associated with low name agreement — that is, self-corrections produced when the number of alternative names is low.

We suggested at the outset that if semantic field size is a significant predictor of the likelihood of prosodic marking, measures of word frequency might be expected to show significant effects too, since both types of measure capture a word's predictability. However, our data provide limited evidence of frequency effects on repair prosody: we only observe a weak effect such that a particularly large decrease in lemma frequency from reparandum item to repair item is associated with a particularly substantial drop in intensity median and mean.

Interestingly, while the effect is weak, its direction would seem to be consistent with Levelt and Cutler's [1983] findings: a decrease in lemma frequency from reparandum to repair item means a decrease in the relative predictability of the repair item; this is on a par with a relatively large semantic field size, so should be associated with a decrease of the likelihood of prosodic marking. Given that in our data, prosodic marking is mostly achieved through an increase on all pitch and intensity delta measures, a decrease of the likelihood of marking corresponds to a decrease on these parameters.

Moreover, our findings regarding the relationship between our frequency variables and other candidate predictors confirm Kapatsinski's [2010] findings on repairs in American English. First, word frequency does not show a systematic relationship with the error–appropriateness distinction: it is not the case that the two semantic subtypes of repair involve different word frequency contours. Second, word frequency does show a systematic relationship with the temporal make-up of the repair: the higher the word or lemma frequency of the reparandum item, the less likely it is to be interrupted prior to repair. This provides support for the notion that higher-frequency lexical items form more cohesive units in speech production [Logan 1982, Bybee 2001, 2002, Kapatsinski 2010].

The temporal make-up of the repairs in turn shows no systematic relationship with repair semantics: unlike Levelt [1989] and Brédart [1991], we do not find that reparanda in error repairs are more likely than reparanda in appropriateness repairs to be interrupted prior to repair. Like our other candidate predictors, repair timing appears to have only a limited effect on repair prosody. Our modelling of the acoustic parameters and prosodic marking judgements revealed no significant effects of any of the temporal variables. The only hint at a systematic relationship emerged in our control procedure, when we modelled the temporal variables using prosodic variables as candidate predictors: in modelling *Proportional completeness*, *Intensity maximum delta* yielded a significant split in the data. The effect is again a weak one, but interestingly, its direction is in line with Nooteboom's [2010] findings on the prosody of phonological error repairs. Nooteboom observes that while repairs in which the interruption comes very early tend to be associated with a high pitch and intensity prominence on the first vowel, repairs in which the erroneous word is completed tend to be associated with a low pitch and intensity prominence. In our data, instances with a large drop in intensity maximum between reparandum and repair are more likely to have a completed reparandum item than instances without such a drop.

The preceding discussion confirms that while we find few significant effects of our predictor variables, insofar as we observe any tendencies in our data, they are consistent with

Levelt and Cutler's [1983] findings on the role of the error–appropriateness distinction and semantic field size in conditioning prosodic marking, and Nooteboom's [2010] observations on the influence of repair timing on pitch and intensity. This may of course be accidental, and we cannot draw firm conclusions from statistically non-significant effects. Still, it is tempting to conclude that the effects described by Levelt and Cutler [1983] and Nooteboom [2010] *do* find some support in our data, but are largely masked by other effects which we have not controlled for in our design, or quantified in our analysis. In particular, we suggested above that in the task-oriented data of Levelt and Cutler [1983], it seems plausible that lexical errors are pragmatically more consequential than appropriateness issues, as the success of the task crucially depends on getting factual instructions right. The relatively high likelihood of prosodic marking of an error repair may be due to this high pragmatic consequentiality of the incorrect information. It also seems plausible that lexical errors are pragmatically more consequential than appropriateness issues in a wide range of discourse contexts. However, as we suggested above, there may well be specific contexts in which appropriateness issues are particularly consequential, and some of these may be represented when repairs are sampled from uncontrolled, spontaneous talk-in-interaction.

A similar argument can be made for the effect of repair timing observed by Nooteboom [2010]: this may emerge as significant when all other things — crucially including the function of the repair in the local discourse context — are equal, and as a weak tendency when they are not. A next step in our research is to investigate the discourse contexts in which the repairs are embedded, to investigate whether there are pragmatic factors that favour or disfavour prosodic marking, which may interact in interesting ways with repair semantics and timing.

CONCLUSION

In this paper we have reported on a study of the prosodic characteristics of lexical self-repair in spontaneous Dutch speech. Our findings confirm the observation, made first by Cutler [1983], that repairs may be produced with or without 'prosodic marking', although the proportion of marked instances in our data is low compared with previous studies, around 20%. We have shown that measures of f0 and intensity maximum and central tendency are strongly correlated with auditory judgements of prosodic marking, with less variation in the implementation of prosodic marking than suggested by Cutler [1983] and Levelt and Cutler [1983]: most 'marked' instances show an increase on both f0 and intensity measures between the reparandum and repair items. With respect to factors conditioning prosodic marking, our

analyses have largely yielded negative results: our data show very few significant effects of the semantic, temporal and frequency-related factors that we might expect to condition repair prosody on the basis of Levelt and Cutler's [1983] and Nooteboom's [2010] findings — although it is perhaps noteworthy that the effects and tendencies that we *do* find are in the expected directions.

ACKNOWLEDGMENTS

This work was supported by ESRC grant RES-061-25-0417 'Prosodic marking revisited: The phonetics of self-initiated self-repair in Dutch'. We thank Christina Englert for her contribution to the research reported here, and the associate editor of *Phonetica* and two anonymous reviewers for helpful comments on the first submission of this paper.

REFERENCES

Baayen, R.H.; Piepenbrock, R.; Gulikers, L.: The CELEX lexical database. Release 2 [CD-ROM]. (Linguistics Data Consortium, Philadelphia 1995)

Wassink, A.B.; Wright, R.A.; Franklin, A.D. Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. J. Phonet. *35:* 363–379 (2007).

Benkenstein, R.; Simpson, A.P.: Phonetic correlates of self-repair involving repetition in German spontaneous speech. Proceedings of DiSS'03 (Disfluency in Spontaneous Speech), Gothenburg 2003.

Boersma, P.; Weenink, D.: Praat: Doing phonetics by computer. Version 5.1.34 (http://www.praat.org/, 2010).

Brédart, S.: Word interruption in self-repairing. Journal of Psycholinguistic Research *20:* 123–137 (1991).

Breiman, L.: Random forests. Machine Learning *45:* 5–32 (2001).

Bybee, J.: Phonology and language use (Cambridge University Press, Cambridge 2001).

Bybee, J.: Word frequency and context of use in the lexical diffusion of phonemically conditioned sound change. Language Variation and Change *14:* 261–290 (2002).

Cole, J.; Hasegawa-Johnson, M.; Shih, C.; Kim, H; Lee, E.-K.; Lu, H.; Mo, Y.; Yoon, T.-J.: Prosodic parallelism as a cue to repetition and error correction disfluency. Proceedings of DiSS'05 (Disfluency in Spontaneous Speech), Aix-en-Provence 2005.

Cutler, A.: Speakers' conceptions of the function of prosody; in Cutler, Ladd, Prosody: Models and measurements, pp. 79–91 (Springer, Heidelberg 1983).

Hartsuiker, R.J.; Notebaert, L.: Lexical access problems lead to disfluencies in speech. Experimental Psychology *57:* 169–177 (2010).

Heemskerk, J.; Zonneveld, W.:. Uitspraakwoordenboek (Het Spectrum, Utrecht 2000).

Hokkanen, T.; Prosodic marking of self-repairs. Proceedings of DiSS'01 (Disfluency in Spontaneous Speech), Edinburgh 2001.

Howell, P.; Young, K.: The use of prosody in highlighting alterations in repairs from unrestricted speech. The Quarterly Journal of Experimental Psychology *43A:* 733–758 (1991).

Jasperson, R.: Some linguistic aspects of closure cut-off; in Ford, Fox, Thompson, The language of turn and sequence, pp. 257–286 (Oxford University Press, Oxford 2002).
Kapatsinski, V.: Frequency of use leads to automaticity of production: Evidence from repair in conversation. Lang. Speech *53:* 71–105 (2010).
Kormos, J.: Monitoring and self-repair in L2. Language Learning *49:* 303–342 (1999).
Levelt, W.J.M.; Cutler, A.: Prosodic marking in speech repair. Journal of Semantics *2:* 205–217 (1983).
Levelt, W.J.M.: Monitoring and self-repair in speech. Cognition *14:* 41–104 (1983).
Local, J.: Variable domains and variable relevance: Interpreting phonetic exponents. J. Phonet. *31:* 321–339 (2003).
Local, J.; Auer, P.; Drew, P.: Retrieving, redoing and resuscitating turns in conversation; in Barth-Weingarten, Reber, Selting, Prosody in interaction, pp. 131–159 (John Benjamins, Amsterdam 2010).
Local, J.; Walker, G.: Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech. Phonetica *62:* 120–130 (2005).
Logan, G.D.: On the ability to inhibit complex movements: A stop-signal study of typewriting. Journal of Experimental Psychology: Human Perception and Performance *8:* 778–792 (1982).
Nakatani, C.H.; Hirschberg, J.: A corpus-based study of repair cues in spontaneous speech. J. Acoust. Soc. Am. *95:* 1603–1616 (1994).
Niebuhr, O.: On the phonetics of intensifying emphasis in German. Phonetica *67:* 1–29 (2010).
Nooteboom, S.: Monitoring for speech errors has different functions in inner and overt speech; in Everaert, Lentz, De Mulder, Nilsen, Zondervan, The linguistic enterprise, pp. 213–233 (John Benjamins, Amsterdam 2010).
Ogden, R.: Phonetics and social action in agreements and disagreements. J. Pragm. *38:* 1752–1775 (2006).
Ogden, R.: Prosodic constructions in making complaints; in Barth-Weingarten, Reber, Selting, Prosody in interaction, pp. 81–103 (John Benjamins, Amsterdam 2010).
Oostdijk, N.: The design of the Spoken Dutch Corpus; in Peters, Collins, Smith, New frontiers of corpus research, pp. 105–113 (Rodopi, Amsterdam 2002).
Plug, L.: Phonetic reduction and informational redundancy in self-initiated self-repair in Dutch. J. Phonet. *39:* 289–297 (2011).
Selting, M.; Prosody in interaction: State of the art; in Barth-Weingarten, Reber, Selting, Prosody in interaction, pp. 3–40 (John Benjamins, Amsterdam 2010).
Severens, E.; van Lommel,S.; Ratinckx, E.; Hartsuiker, R.J.: Timed picture naming norms for 590 pictures in Dutch. Acta Psychologica *119:* 159–187 (2005).
Shriberg, E.: To 'errr' is human: Ecology and acoustics of speech disfluencies. J.Int. Phonet. Assoc. *31:* 153–169 (2001).
Sluijter, A.M.C.; van Heuven, V.J.: Spectral balance as an acoustic correlate of linguistic stress. J. Acoust. Soc. Am. *100:* 2471–2485 (1994).
Smiljanić, R.; Bradlow, A.R.: Speaking and hearing clearly: Talker and listener factors in speaking style changes. Language and Linguistics Compass *3:* 236–264 (2009).
Strobl, C.; Malley, J.; Tutz, G.: An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods *14:* 323–348 (2009).
Tagliamonte, S.; Baayen, R.H.: Models, forests and trees of York English: Was/were variation as a case study for statistical practice. Language Variation and Change *24:* 135–178 (2012).
Walker, T.: The phonetics of sequence organization: An investigation of lexical repetition in other-initiated repair sequences in American English (VDM Verlag Dr. Mueller, Saarbrücken 2009).